Geometric Analysis of Hilbert-Schmidt Independence Criterion Based ICA Contrast Function

NICTA Technical Report: PA006080 (ISSN 1833-9646)

Hao Shen[1], Stefanie Jegelka[2] and Arthur Gretton[2]

[1] Systems Engineering and Complex Systems Research Program,
National ICT Australia **, Canberra ACT 2601, Australia
and
Department of Information Engineering,
Research School of Information Science and Engineering,
The Australian National University, Canberra ACT 0200, Australia

[2] Department of Empirical Inference for Machine Learning and Perception,
Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany

Hao.Shen@rsise.anu.edu.au, {Stefanie.Jegelka,Arthur.Gretton}@tuebingen.mpg.de

1 Introduction

Since the success of Independent Component Analysis (ICA) for solving the Blind Source Separation (BSS) problem [1, 2], ICA has received considerable attention in numerous areas, such as signal processing, statistical modeling, and unsupervised learning. The performance of ICA algorithms depends significantly on the choice of the contrast function measuring statistical independence of signals and on the appropriate optimisation technique.

From an independence criterion's point of view, there exist numerous parametric and nonparametric approaches for designing ICA contrast functions. It has been well known that parametric ICA methods are rather limited to particular families of sources [3]. For these parametric approaches, contrast functions are selected according to certain hypothetical distributions (probability density functions) of the sources by a single fixed nonlinear function. In practical applications, however, the distributions of the sources are unknown, and even can not be approximated by a single function. Therefore parametric ICA methods have their fatal weakness in handling many real applications.

It is well known that nonparametric methods have their capability and robustness of estimating unknown distributions of the sources. Recently there have been many interests in designing nonparametric ICA contrast function. One of the possibilities is to use kernel density estimation to deal with the unknown source distributions, such as [4, 5]. There also exist other nonparametric ICA methods, which do not work with the probability density estimator directly, such as [6–8]. Most recently, the so-called Hilbert-Schmidt Independence Criterion (HSIC) was proposed for measuring statistical independence between two random variables [9]. In the sequel, an HSIC based ICA contrast has

** National ICT Australia is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through *Backing Australia's Ability* and the ICT Research Centre of Excellence programs.

shown its superiority over most of the other nonparametric approaches in terms of high quality separations. Although Jegelka and Gretton [10] have developed a gradient descent based method with a local quadratic step search strategy to optimise this HSIC based ICA contrast, the question of better search direction choice, which could provide high order convergence, has not been addressed.

On the other hand, from an optimisation's point view, since the influential paper of Comon [1], many efficient ICA algorithms have been developed by researchers from various communities. Recently, there has been an increasing interest in using geometric optimisation for ICA problems. Using the geometric optimisation techniques, the first author and his colleagues have developed a big family of algorithms by means of approximate Newton/Newton-like methods on proper manifold settings [11–14]. As an aside, a very popular one-unit ICA method [15], the FastICA algorithm, can be just considered as a special case in this family. One crucial technical insight of these methods is the structure of the Hessians at a correct separation point. It deduces a sensible and efficient approximation of the Hessians at an arbitrary point within a neighborhood of a correct separation point and consequently leads to approximate Newton-like ICA methods which ensure local quadratic convergence.

Besides the gradient-based optimisation techniques above, a family of Jacobi-type algorithms is also commonly used in the ICA community [16, 1, 17, 18]. Although numerical evidences have shown the effectiveness and efficiency of these methods, the stability and convergence properties are still theoretically unknown for general settings. According to recent results in the convergence analysis of general Jacobi-type algorithms [19–21], it has been shown that the convergence properties of Jacobi-type methods depend significantly on search directions with respect to the structure of the Hessian of the cost function being optimised. According to our survey, a Jacobi-type method for optimising the HSIC based ICA contrast has actually already been proposed in [7], which was however in a completely different context. Nevertheless the convergence properties have not been addressed yet. Therefore the motivation of this work is to explore the Hessian structure of the HSIC based ICA contrast to form a basis for future developments and analysis of both approximate Newton-like methods and Jacobi-type methods for optimising the HSIC based ICA contrast.

Moreover, a generalisation of HSIC for measuring mutual statistical independence between more than two random variables has already been proposed by Kankainen in [22]. It led to the so-called characteristic-function-based ICA contrast function (CFICA) [7], where HSIC can be just considered as a pairwise case. This contrast function has been tackled by a gradient descent method on the special orthogonal group with a golden search [8]. However this approach generally suffers with enormous computational burden. Hence to investigate the feasibility of using either the approximate Newton-like method or Jacobi-type method for optimising the CFICA with better convergence properties, we analyse the CFICA contrast in the same fashion as the HSIC based ICA contrast.

The report is organised as follows. In Section 2, we review the HSIC based ICA contrast. Section 3 characterises the critical point condition of HSIC based ICA contrast. It turns out that any correct separation matrix is a critical point of the HSIC based ICA contrast. Moreover, our analysis shows that the Hessian of the HSIC based ICA contrast at a correct separation matrix is indeed diagonal. In Section 4, the CFICA constrast function is analysed in the same fashion as the HSIC based ICA. It shows that any correct separation matrix fulfills the critical point condition of CFICA, whereas the Hessian of CFICA at a correct separation matrix is generally not diagonal. Finally, a brief conclusion and suggestions for future work are given in Section 5.

2 Problem Setting

We consider the standard noiseless instantaneous whitened linear ICA demixing model [2]

$$Y = X^{\top} W, \tag{1}$$

where $W := [w_1, \ldots, w_n] \in \mathbb{R}^{m \times n}$ with $m \ll n$ is the whitened observation, the orthogonal matrix $X := [x_1, \ldots, x_m] \in \mathbb{R}^{m \times m}$, i.e., $X^{\top} X = I$, is the demixing matrix, and $Y \in \mathbb{R}^{m \times n}$ is the recovered signal. Let denote the special orthogonal group $SO(m) := \{X \in \mathbb{R}^{m \times m} | X^{\top} X = I, \det X = 1\}$. Without loss of generality, we restrict the demixing matrix $X \in SO(m)$.

Originally, HSIC only measures the statistical independence between two random variables, i.e., pairwisely. Let $u, v \in \mathbb{R}$ be two real valued random variables. HSIC can be formulated as follows, see [9] for details,

$$\begin{aligned} HSIC(p_{u,v}, \mathcal{F}, \mathcal{G}) = {} & \mathbb{E}_{u,u',v,v'} \left[\phi\left(u, u'\right) \psi\left(v, v'\right) \right] + \mathbb{E}_{u,u'} \left[\phi\left(u, u'\right) \right] \mathbb{E}_{v,v'} \left[\psi\left(v, v'\right) \right] \\ & - 2 \mathbb{E}_{u,v} \left[\mathbb{E}_{u'} \left[\phi\left(u, u'\right) \right] \mathbb{E}_{v'} \left[\psi\left(v, v'\right) \right] \right], \end{aligned} \tag{2}$$

where $\mathcal{F}$ and $\mathcal{G}$ are two Hilbert spaces of functions from a compact subset $\mathbb{R} \supset \mathcal{U}$ to $\mathbb{R}$, $\phi(\cdot)$ and $\psi(\cdot)$ are certain kernel functions, and $\mathbb{E}_{u,u',v,v'}[\cdot]$ denotes the expectation over independent identical pairs u, v and u', v'. Note that ϕ and ψ are not necessarily different. We specify the Gaussian kernel function as a concrete example

$$\phi(t_1, t_2) = \phi(t_1 - t_2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(t_1 - t_2)^2}{2h^2} \right), \qquad \text{where } t_1, t_2 \in \mathbb{R}. \tag{3}$$

To simplify the complexity of analysis, we assume $h = 1$. Moreover in the ICA setting as in (1), assume that u and v are whitened, then the following results hold true

$$\begin{cases} \mathbb{E}_u[u] & = 0 \\ \mathbb{E}_u[u^2] & = 1 \end{cases} \qquad \text{and} \qquad \begin{cases} \mathbb{E}_{u,v}[u - v] & = 0 \\ \mathbb{E}_{u,v}[(u - v)^2] & = 2. \end{cases} \tag{4}$$

Now let us consider ICA problems with $m > 2$ signals. It has been shown that in the ICA setting, the mutual independence between all signals can be ensured by the pairwise independences [1]. Hence recall the instantaneous ICA demixing model as in (1). By summing up all unique pairwise HSICs, the overall HSIC score over the estimated signals $Y \in \mathbb{R}^{m \times n}$ can be computed as follows

$$\begin{aligned} H : SO(m) \to \mathbb{R}, \qquad H(X) := \sum_{1 \le i < j \le m}^{m} & \mathbb{E}_{k,l} \left[\phi\left(x_i^{\top} \overline{w}_{kl}\right) \phi\left(x_j^{\top} \overline{w}_{kl}\right) \right] \\ & + \mathbb{E}_{k,l} \left[\phi\left(x_i^{\top} \overline{w}_{kl}\right) \right] \mathbb{E}_{k,l} \left[\phi\left(x_j^{\top} \overline{w}_{kl}\right) \right] \\ & - 2 \mathbb{E}_{k} \left[\mathbb{E}_{l} \left[\phi\left(x_i^{\top} \overline{w}_{kl}\right) \right] \mathbb{E}_{l} \left[\phi\left(x_j^{\top} \overline{w}_{kl}\right) \right] \right], \end{aligned} \tag{5}$$

where $\overline{w}_{kl} = w_k - w_l \in \mathbb{R}^m$ denotes the difference between k-th and l-th sample instances, and $\mathbb{E}_{k,l}[\cdot]$ represents the empirical expectation over sample indices k and l.

3 Geometric Analysis of HSIC

3.1 Critical Point Analysis of HSIC

In this section we show that a demixing matrix, which corresponds to correct separations, can be attained at a certain optimum of HSIC, i.e., the correct demixing matrix fulfills the critical point condition of HSIC.

Let $\mathfrak{so}(m) := \{\Omega \in \mathbb{R}^{m\times m} | \Omega = -\Omega^\top\}$ denote the set of all $m \times m$ skew-symmetric matrices. Recall the geodesic of $SO(m)$ emanating from a point $X \in SO(m)$ as

$$\gamma_X : \mathbb{R} \to SO(m), \qquad \varepsilon \mapsto X \exp(\varepsilon X^\top \Xi), \tag{6}$$

where $X = [x_1, \ldots, x_m] \in SO(m)$ and $\Xi = [\xi_1, \ldots, \xi_m] \in T_X SO(m)$. Here $T_X SO(m)$ denotes the tangent space of $SO(m)$ at a point X, i.e.,

$$T_X SO(m) := \left\{\Xi \in \mathbb{R}^{m\times m} | \Xi = X\Omega,\ \Omega \in \mathfrak{so}(m)\right\}. \tag{7}$$

Now by the chain rule, the first derivative of H can be computed as follows

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\,\varepsilon}(H \circ \gamma_X)(\varepsilon)\bigg|_{\varepsilon=0} = \sum_{i,j=1, i\neq j}^{m} & \mathbb{E}_{k,l}\left[\phi'\left(x_i^\top \overline{w}_{kl}\right)\xi_i^\top \overline{w}_{kl}\phi\left(x_j^\top \overline{w}_{kl}\right)\right] \\ & + \mathbb{E}_{k,l}\left[\phi'\left(x_i^\top \overline{w}_{kl}\right)\xi_i^\top \overline{w}_{kl}\right]\mathbb{E}_{k,l}\left[\phi\left(x_j^\top \overline{w}_{kl}\right)\right] \\ & - 2\mathbb{E}_k\left[\mathbb{E}_l\left[\phi'\left(x_i^\top \overline{w}_{kl}\right)\xi_i^\top \overline{w}_{kl}\right]\mathbb{E}_l\left[\phi\left(x_j^\top \overline{w}_{kl}\right)\right]\right]. \end{aligned} \tag{8}$$

Set the first derivative of H as in (8) equal to zero, one can characterise critical points of the HSIC based ICA contrast function as in (5). It is obvious that such a critical point condition depends not only on the statistical characteristics of sources, but also on the properties of kernel functions. It is hardly possible to characterise all critical points of HSIC in full generality.

Nevertheless, we will show that a correct demixing matrix $X^* \in SO(m)$ is indeed a critical point of H. Let $S := [s_1, \ldots, s_n] = X^{*\top} W \in \mathbb{R}^{m\times n}$ denote the recovered statistically independent components, and let $\Omega := [\omega_1, \ldots, \omega_m] = (\omega_{ij})_{i,j=1}^m \in \mathfrak{so}(m)$. By evaluating the demixing model (1) at a correct separation matrix $X^* \in SO(m)$, one easily gets

$$\begin{cases} x_i^{*\top}\overline{w}_{kl} = \overline{s}_{kli} \\ \xi_i^{*\top}\overline{w}_{kl} = \omega_i^\top \overline{s}_{kl} \end{cases}, \tag{9}$$

where $\overline{s}_{kl} := s_k - s_l \in \mathbb{R}^m$ and $\overline{s}_{kli} = e_i^\top \overline{s}_{kl} \in \mathbb{R}$ is the i-th entry of $\overline{s}_{kl}$. Hence the evaluation of (8) at X^* can be computed as follows

$$\frac{\mathrm{d}}{\mathrm{d}\,\varepsilon}(H \circ \gamma_{X^*})(\varepsilon)\bigg|_{\varepsilon=0} = \sum_{i,j=1, i\neq j}^{m} \mathbb{E}_{k,l}\left[\phi'\left(\overline{s}_{kli}\right)\omega_i^\top \overline{s}_{kl}\phi\left(\overline{s}_{klj}\right)\right] + \mathbb{E}_{k,l}\left[\phi'\left(\overline{s}_{kli}\right)\omega_i^\top \overline{s}_{kl}\right]\mathbb{E}_{k,l}\left[\phi\left(\overline{s}_{klj}\right)\right] \tag{10a}$$

$$- 2\mathbb{E}_k\left[\mathbb{E}_l\left[\phi'\left(\overline{s}_{kli}\right)\omega_i^\top \overline{s}_{kl}\right]\mathbb{E}_l\left[\phi\left(\overline{s}_{klj}\right)\right]\right] \tag{10b}$$

$$= \sum_{i,j,r=1, i\neq j,r}^{m} \mathbb{E}_{k,l}\left[\phi'\left(\overline{s}_{kli}\right)\omega_{ir}\overline{s}_{klr}\phi\left(\overline{s}_{klj}\right)\right] + \mathbb{E}_{k,l}\left[\phi'\left(\overline{s}_{kli}\right)\omega_{ir}\overline{s}_{klr}\right]\mathbb{E}_{k,l}\left[\phi\left(\overline{s}_{klj}\right)\right] \tag{10c}$$

$$- 2\mathbb{E}_k\left[\mathbb{E}_l\left[\phi'\left(\overline{s}_{kli}\right)\omega_{ir}\overline{s}_{klr}\right]\mathbb{E}_l\left[\phi\left(\overline{s}_{klj}\right)\right]\right] \tag{10d}$$

The properties of the prewhitened data (4) imply that the second term in (10c) vanishes, and the above equation (10) can be simplified as

$$\frac{\mathrm{d}}{\mathrm{d}\,\varepsilon}(H \circ \gamma_{X^*})(\varepsilon)\Big|_{\varepsilon=0} = \sum_{i,j=1,i\neq j}^{m} \omega_{ij}\mathbb{E}_{k,l}\left[\phi'\left(\overline{s}_{kli}\right)\right]\left(\mathbb{E}_{k,l}\left[\overline{s}_{klj}\phi\left(\overline{s}_{klj}\right)\right] - 2\mathbb{E}_{k}\left[\overline{s}_{klj}\mathbb{E}_{l}\left[\phi\left(\overline{s}_{klj}\right)\right]\right]\right). \tag{11}$$

Finally, applying the symmetry of the kernel function, i.e., $\mathbb{E}_{k,l}\left[\phi'\left(\overline{s}_{kli}\right)\right] = 0$, one can conclude that the first derivative of H as in (8) vanishes at a correct separation point X^*, i.e., any correct separation matrix is a critical point of the HSIC contrast function H as in (5).

3.2 Exploration of the Structure of the Hessian of HSIC

In this section, we explore the structure of the Hessian of HSIC contrast H as in (5) at a correct separation X^*. Now take the second derivative of H, one gets

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}\,\varepsilon^2}(H \circ \gamma_{X})(\varepsilon)\Big|_{\varepsilon=0} = \sum_{i,j=1,i\neq j}^{m} & \mathbb{E}_{k,l}\left[\phi''\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top\overline{w}_{kl}\overline{w}_{kl}^\top\xi_i\phi\left(x_j^\top\overline{w}_{kl}\right)\right] && \text{(12a)}\\
& -\mathbb{E}_{k,l}\left[\phi'\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top \varXi X^\top\overline{w}_{kl}\phi\left(x_j^\top\overline{w}_{kl}\right)\right] && \text{(12b)}\\
& +\mathbb{E}_{k,l}\left[\phi'\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top\overline{w}_{kl}\overline{w}_{kl}^\top\xi_j\phi'\left(x_j^\top\overline{w}_{kl}\right)\right] && \text{(12c)}\\
& +\mathbb{E}_{k,l}\left[\phi''\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top\overline{w}_{kl}\overline{w}_{kl}^\top\xi_i\right]\mathbb{E}_{k,l}\left[\phi\left(x_j^\top\overline{w}_{kl}\right)\right] && \text{(12d)}\\
& -\mathbb{E}_{k,l}\left[\phi'\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top \varXi X^\top\overline{w}_{kl}\right]\mathbb{E}_{k,l}\left[\phi\left(x_j^\top\overline{w}_{kl}\right)\right] && \text{(12e)}\\
& +\mathbb{E}_{k,l}\left[\phi'\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top\overline{w}_{kl}\right]\mathbb{E}_{k,l}\left[\phi'\left(x_j^\top\overline{w}_{kl}\right)\xi_j^\top\overline{w}_{kl}\right] && \text{(12f)}\\
& -2\mathbb{E}_{k}\left[\mathbb{E}_{l}\left[\phi''\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top\overline{w}_{kl}\overline{w}_{kl}^\top\xi_i\right]\mathbb{E}_{l}\left[\phi\left(x_j^\top\overline{w}_{kl}\right)\right]\right] && \text{(12g)}\\
& +2\mathbb{E}_{k}\left[\mathbb{E}_{l}\left[\phi'\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top \varXi X^\top\overline{w}_{kl}\right]\mathbb{E}_{l}\left[\phi\left(x_j^\top\overline{w}_{kl}\right)\right]\right] && \text{(12h)}\\
& -2\mathbb{E}_{k}\left[\mathbb{E}_{l}\left[\phi'\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top\overline{w}_{kl}\right]\mathbb{E}_{l}\left[\phi'\left(x_j^\top\overline{w}_{kl}\right)\xi_j^\top\overline{w}_{kl}\right]\right]. && \text{(12i)}
\end{aligned}$$

Let $X = X^*$. The first term (12a) can be computed as

$$\mathbb{E}_{k,l}\left[\phi''\left(\overline{s}_{kli}\right)\omega_i^\top\overline{s}_{kl}\overline{s}_{kl}^\top\omega_i\phi\left(\overline{s}_{klj}\right)\right] = \sum_{r,t=1,r,t\neq i}^{m}\omega_{ir}\omega_{it}\mathbb{E}_{k,l}\left[\phi''\left(\overline{s}_{kli}\right)\overline{s}_{klr}\overline{s}_{klt}\phi\left(\overline{s}_{klj}\right)\right]. \tag{13}$$

The properties of the whitened data, (4), imply $\mathbb{E}_{k,l}\left[\phi''\left(\overline{s}_{kli}\right)\overline{s}_{klr}\overline{s}_{klt}\phi\left(\overline{s}_{klj}\right)\right] = 0$ for all $r \neq t$ and $r, t \neq i$. Thus one further gets

$$\begin{aligned}
\mathbb{E}_{k,l}\left[\phi''\left(\overline{s}_{kli}\right)\omega_i^\top\overline{s}_{kl}\overline{s}_{kl}^\top\omega_i\phi\left(\overline{s}_{klj}\right)\right] = & \sum_{r=1,r\neq i,j}^{m} 2\omega_{ir}^2\mathbb{E}_{k,l}[\phi''\left(\overline{s}_{kli}\right)]\mathbb{E}_{k,l}[\phi\left(\overline{s}_{klj}\right)]\\
& + \omega_{ij}^2\mathbb{E}_{k,l}[\phi''\left(\overline{s}_{kli}\right)]\mathbb{E}_{k,l}\left[\overline{s}_{klj}^2\phi\left(\overline{s}_{klj}\right)\right].
\end{aligned} \tag{14}$$

Hence by applying the exactly same techniques, the remaining terms (12b)-(12i) can be computed as follows,

$$\text{(12b)}:\ \mathbb{E}_{k,l}\left[\phi'\left(\overline{s}_{kli}\right)\omega_i^\top \varOmega\overline{s}_{kl}\phi\left(\overline{s}_{klj}\right)\right] = \sum_{r=1,r\neq i}^{m} -\omega_{ir}^2\mathbb{E}_{k,l}\left[\phi'\left(\overline{s}_{kli}\right)\overline{s}_{kli}\right]\mathbb{E}_{k,l}\left[\phi\left(\overline{s}_{klj}\right)\right], \tag{15}$$

$$\text{(12c)}: \ \mathbb{E}_{k,l}\left[\phi'\left(\bar{s}_{kli}\right)\omega_i^\top \bar{s}_{kl}\bar{s}_{kl}^\top \omega_j \phi'\left(\bar{s}_{klj}\right)\right] = -\omega_{ij}^2 \mathbb{E}_{k,l}\left[\phi'\left(\bar{s}_{kli}\right)\bar{s}_{kli}\right]\mathbb{E}_{k,l}\left[\phi'\left(\bar{s}_{klj}\right)\bar{s}_{klj}\right], \tag{16}$$

$$\text{(12d)}: \ \mathbb{E}_{k,l}\left[\phi''\left(\bar{s}_{kli}\right)\omega_i^\top \bar{s}_{kl}\bar{s}_{kl}^\top \omega_i\right]\mathbb{E}_{k,l}\left[\phi\left(\bar{s}_{klj}\right)\right] = \sum_{r=1, r\neq i}^{m} 2\omega_{ir}^2 \mathbb{E}_{k,l}\left[\phi''\left(\bar{s}_{kli}\right)\right]\mathbb{E}_{k,l}\left[\phi\left(\bar{s}_{klj}\right)\right], \tag{17}$$

$$\text{(12e)}: \ \mathbb{E}_{k,l}\left[\phi'\left(\bar{s}_{kli}\right)\omega_i^\top \Omega \bar{s}_{kl}\right]\mathbb{E}_{k,l}\left[\phi\left(\bar{s}_{klj}\right)\right] = \sum_{r=1, r\neq i}^{m} -\omega_{ir}^2 \mathbb{E}_{k,l}\left[\phi'\left(\bar{s}_{kli}\right)\bar{s}_{kli}\right]\mathbb{E}_{k,l}\left[\phi\left(\bar{s}_{klj}\right)\right], \tag{18}$$

$$\text{(12f)}: \ \mathbb{E}_{k,l}\left[\phi'\left(\bar{s}_{kli}\right)\omega_i^\top \bar{s}_{kl}\right]\mathbb{E}_{k,l}\left[\phi'\left(\bar{s}_{klj}\right)\omega_j^\top \bar{s}_{kl}\right] = 0, \tag{19}$$

$$\begin{aligned}\text{(12g)}: \ \mathbb{E}_{k}\left[\mathbb{E}_{l}\left[\phi''\left(\bar{s}_{kli}\right)\omega_i^\top \bar{s}_{kl}\bar{s}_{kl}^\top \omega_i\right]\mathbb{E}_{l}\left[\phi\left(\bar{s}_{klj}\right)\right]\right] &= \omega_{ij}^2 \mathbb{E}_{k,l}\left[\phi''\left(\bar{s}_{kli}\right)\right]\mathbb{E}_{k}\left[\mathbb{E}_{l}\left[\bar{s}_{klj}^2\right]\mathbb{E}_{l}\left[\phi\left(\bar{s}_{klj}\right)\right]\right] \\ &\quad + \sum_{r=1, r\neq i,j}^{m} 2\omega_{ir}^2 \mathbb{E}_{k,l}\left[\phi''\left(\bar{s}_{kli}\right)\right]\mathbb{E}_{k,l}\left[\phi\left(\bar{s}_{klj}\right)\right],\end{aligned} \tag{20}$$

$$\text{(12h)}: \ \mathbb{E}_{k}\left[\mathbb{E}_{l}\left[\phi'\left(\bar{s}_{kli}\right)\omega_i^\top \Omega \bar{s}_{kl}\right]\mathbb{E}_{l}\left[\phi\left(\bar{s}_{klj}\right)\right]\right] = \sum_{r=1, r\neq i}^{m} -\omega_{ir}^2 \mathbb{E}_{k,l}\left[\phi'\left(\bar{s}_{kli}\right)\bar{s}_{kli}\right]\mathbb{E}_{k,l}\left[\phi\left(\bar{s}_{klj}\right)\right], \tag{21}$$

$$\text{(12i)}: \mathbb{E}_{k}\left[\mathbb{E}_{l}\left[\phi'(\bar{s}_{kli})\,\omega_i^\top \bar{s}_{kl}\right]\mathbb{E}_{l}\left[\phi'(\bar{s}_{klj})\,\omega_j^\top \bar{s}_{kl}\right]\right] = -\omega_{ij}^2 \mathbb{E}_{k}[\mathbb{E}_{l}[\bar{s}_{kli}]\mathbb{E}_{l}[\phi'(\bar{s}_{kli})]]\mathbb{E}_{k}[\mathbb{E}_{l}[\bar{s}_{klj}]\mathbb{E}_{l}[\phi'(\bar{s}_{klj})]]. \tag{22}$$

Now let us substitute the above results (14)–(22) into (12), the first derivative of H evaluated at X^* can be computed as

$$\left.\frac{\mathrm{d}^2}{\mathrm{d}\,\varepsilon^2}(H \circ \gamma_{X^*})(\varepsilon)\right|_{\varepsilon=0} = \sum_{i,j=1, i\neq j}^{m} \omega_{ij}^2 \kappa_{ij} = \sum_{1\leq i<j\leq m}^{m} \omega_{ij}^2\left(\kappa_{ij} + \kappa_{ji}\right), \tag{23}$$

where

$$\begin{aligned}\kappa_{ij} &= \mathbb{E}_{k,l}[\phi''(\bar{s}_{kli})]\mathbb{E}_{k,l}[(\bar{s}_{klj})^2\phi(\bar{s}_{klj})] + 2\mathbb{E}_{k,l}[\phi''(\bar{s}_{kli})]\mathbb{E}_{k,l}[\phi(\bar{s}_{klj})] \\ &\quad - 2\mathbb{E}_{k,l}[\phi''(\bar{s}_{kli})]\mathbb{E}_{k}\left[\mathbb{E}_{l}[(\bar{s}_{klj})^2]\mathbb{E}_{l}[\phi(\bar{s}_{klj})]\right] - \mathbb{E}_{k,l}[\phi'(\bar{s}_{kli})\bar{s}_{kli}]\mathbb{E}_{k,l}[\phi'(\bar{s}_{klj})\bar{s}_{klj}] \\ &\quad + 2\mathbb{E}_{k}\left[\mathbb{E}_{l}[\bar{s}_{kli}]\mathbb{E}_{l}[\phi'(\bar{s}_{kli})]\right]\mathbb{E}_{k}\left[\mathbb{E}_{l}[\bar{s}_{klj}]\mathbb{E}_{l}[\phi'(\bar{s}_{klj})]\right].\end{aligned} \tag{24}$$

Without loss of generality, let $\Omega = (\omega_{ij})_{i,j=1}^m \in \mathfrak{so}(m)$ and let $\overline{\Omega} = (\omega_{ij})_{1\leq i<j\leq m} \in \mathbb{R}^{m(m-1)/2}$ in a lexicographic order. Obviously, the quadratic form (23) is simply a sum of pure squares, which indicates that the Hessian of the contrast function H in (5) at the desired critical point $X^* \in O(m)$, i.e. the symmetric bilinear form $\mathcal{H}H(X^*)\colon T_{X^*}O(m) \times T_{X^*}O(m) \to \mathbb{R}$, is indeed *diagonal* with respect to the standard basis of $\mathbb{R}^{m(m-1)/2}$.

4 Geometric Analysis of Characteristic-Function-Based ICA

Now in this section we examine the CFICA contrast function in the same fashion as above. As a generalisation of HSIC, the CFICA contrast function [7] is defined as follows

$$F : SO(m) \to \mathbb{R}, \qquad F(X) := \prod_{j=1}^{m} \mathbb{E}_{k,l}\left[\phi\left(x_j^\top \overline{w}_{kl}\right)\right] - 2\mathbb{E}_{k}\left[\prod_{j=1}^{m} \mathbb{E}_{l}\left[\phi\left(x_j^\top \overline{w}_{kl}\right)\right]\right]. \tag{25}$$

Take the first derivative of F, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\,\varepsilon}(F\circ\gamma_X)(\varepsilon)\Big|_{\varepsilon=0} =& \sum_{i=1}^{m}\mathbb{E}_{k,l}\left[\phi'\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top\overline{w}_{kl}\right]\prod_{j=1,j\neq i}^{m}\mathbb{E}_{k,l}\left[\phi\left(x_j^\top\overline{w}_{kl}\right)\right]\\ &-2\sum_{i=1}^{m}\mathbb{E}_k\left[\mathbb{E}_l\left[\phi'\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top\overline{w}_{kl}\right]\prod_{j=1,j\neq i}^{m}\mathbb{E}_l\left[\phi\left(x_j^\top\overline{w}_{kl}\right)\right]\right].\end{aligned} \tag{26}$$

Again, it still seems to be hardly possible to characterise all critical points of F. Nevertheless, by the same arguments as before, one can still show that at a correct separation point $X^*\in SO(m)$, the expression in (26) vanishes, i.e., any correct separation matrix is a critical point of the characteristic contrast F as in (25).

Now take the second derivative of F, one gets

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\,\varepsilon^2}(F\circ\gamma_X)(\varepsilon)\Big|_{\varepsilon=0} =& \sum_{i=1}^{m}\mathbb{E}_{k,l}\left[\phi''\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top\overline{w}_{kl}\overline{w}_{kl}^\top\xi_i\right]\prod_{j=1,j\neq i}^{m}\mathbb{E}_{k,l}\left[\phi\left(x_j^\top\overline{w}_{kl}\right)\right] && \text{(27a)}\\ &-\sum_{i=1}^{m}\mathbb{E}_{k,l}\left[\phi'\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top\varXi X^\top\overline{w}_{kl}\right]\prod_{j=1,j\neq i}^{m}\mathbb{E}_{k,l}\left[\phi\left(x_j^\top\overline{w}_{kl}\right)\right] && \text{(27b)}\\ &+\sum_{h,i=1,h\neq i}^{m}\prod_{j}^{\{h,i\}}\mathbb{E}_{k,l}\left[\phi'\left(x_j^\top\overline{w}_{kl}\right)\xi_j^\top\overline{w}_{kl}\right]\prod_{j=1,j\neq h,i}^{m}\mathbb{E}_{k,l}\left[\phi\left(x_j^\top\overline{w}_{kl}\right)\right] && \text{(27c)}\\ &-2\sum_{i=1}^{m}\mathbb{E}_k\left[\mathbb{E}_l\left[\phi''\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top\overline{w}_{kl}\overline{w}_{kl}^\top\xi_i\right]\prod_{j=1,j\neq i}^{m}\mathbb{E}_l\left[\phi\left(x_j^\top\overline{w}_{kl}\right)\right]\right] && \text{(27d)}\\ &+2\sum_{i=1}^{m}\mathbb{E}_k\left[\mathbb{E}_l\left[\phi'\left(x_i^\top\overline{w}_{kl}\right)\xi_i^\top\varXi X^\top\overline{w}_{kl}\right]\prod_{j=1,j\neq i}^{m}\mathbb{E}_l\left[\phi\left(x_j^\top\overline{w}_{kl}\right)\right]\right] && \text{(27e)}\\ &-2\sum_{h,i=1,h\neq i}^{m}\mathbb{E}_k\left[\prod_{j}^{\{h,i\}}\mathbb{E}_l\left[\phi'\left(x_j^\top\overline{w}_{kl}\right)\xi_j^\top\overline{w}_{kl}\right]\prod_{j=1,j\neq h,i}^{m}\mathbb{E}_l\left[\phi\left(x_j^\top\overline{w}_{kl}\right)\right]\right]. && \text{(27f)}\end{aligned}$$

Let $X=X^*$. The first term (27a) in the right-hand side of (27) is computed as

$$\begin{aligned}&\sum_{i=1}^{m}\mathbb{E}_{k,l}\left[\phi''\left(\overline{s}_{kli}\right)\omega_i^\top\overline{s}_{kl}\overline{s}_{kl}^\top\omega_i\right]\prod_{j=1,j\neq i}^{m}\mathbb{E}_{k,l}\left[\phi\left(\overline{s}_{klj}\right)\right]\\ =&\sum_{i,r=1,r\neq i}^{m}2\,\omega_{ir}^2\mathbb{E}_{k,l}\left[\phi''\left(\overline{s}_{kli}\right)\right]\prod_{j=1,j\neq i}^{m}\mathbb{E}_{k,l}\left[\phi\left(\overline{s}_{klj}\right)\right].\end{aligned} \tag{28}$$

An even more tedious computation evaluates the expression (27d) at X^* as follows

$$\begin{aligned}
&\sum_{i=1}^{m}\mathbb{E}_k\left[\mathbb{E}_l\left[\phi''\left(\overline{s}_{kli}\right)\omega_i^\top\overline{s}_{kl}\overline{s}_{kl}^\top\omega_i\right]\prod_{j=1,j\neq i}^{m}\mathbb{E}_l\left[\phi\left(\overline{s}_{klj}\right)\right]\right]\\
&=\sum_{i,r=1,r\neq i}^{m}\omega_{ir}^2\mathbb{E}_{k,l}\left[\phi''\left(\overline{s}_{kli}\right)\right]\mathbb{E}_k\left[\mathbb{E}_l\left[\overline{s}_{klr}^2\right]\mathbb{E}_l\left[\phi\left(\overline{s}_{klr}\right)\right]\right]\prod_{j=1,j\neq i,r}^{m}\mathbb{E}_{k,l}\left[\phi\left(\overline{s}_{klj}\right)\right]\\
&\quad+\sum_{i,r,t=1,r\neq t,r,t\neq i}^{m}\omega_{ir}\omega_{it}\mathbb{E}_{k,l}\left[\phi''\left(\overline{s}_{kli}\right)\right]\prod_{j=1}^{\{r,t\}}\mathbb{E}_k\left[\mathbb{E}_l\left[\overline{s}_{klj}\right]\mathbb{E}_l\left[\phi\left(\overline{s}_{klj}\right)\right]\right]\prod_{j=1,j\neq i,r,t}^{m}\mathbb{E}_{k,l}\left[\phi\left(\overline{s}_{klj}\right)\right].
\end{aligned}\tag{29}$$

Thus, the second summand on the right-hand side of (29) is clearly not a sum of pure squares. Following the same reasoning, an even more tedious computation turns out that the Hessian of H (5) at a correct separation point $X^* \in SO(m)$ is a dense matrix, in general. In other words, let $\Omega = (\omega_{ij})_{i,j=1}^m \in \mathfrak{so}(m)$ and let $\overline{\Omega} = (\omega_{ij})_{1\leq i<j\leq m} \in \mathbb{R}^{m(m-1)/2}$ in a lexicographic order, the Hessian of H at $X^* \in SO(m)$ is neither diagonal with respect to the standard basis of $\mathbb{R}^{m(m-1)/2}$, nor enjoys another simple structure.

5 Conclusions and Future Work

In this work, we rigorously derive the critical conditions of both the HSIC based ICA contrast and the CFICA contrast. It turns out that any correct separation matrix is indeed a critical point of both contrast functions. Further analysis shows that the Hessian of the HSIC based contrast is **diagonal** at a correct separation point, whereas this does not generally hold true for CFICA. Therefore our future work will focus on the development and analysis of both approximate Newton-like methods and Jacobi-type methods for optimising the HSIC based ICA contrast rather that the CFICA contrast.

Acknowledgment

The first author wishes to thank the Max Planck Institute for Biological Cybernetics for the hospitality of supporting and inviting him to work with Ms. Stefanie Jegelka and Dr. Arthur Gretton.

This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

1. P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
2. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley, New York, 2001.
3. J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proceedings of the IEEE, Special Issue on Blind Identification and Estimation*, vol. 86, no. 10, pp. 2009–2025, 1998.
4. R. Boscolo, H. Pan, and V. P. Roychowdhury, "Independent component analysis based on nonparametric density estimation," *IEEE Transactions on Neural Networks*, vol. 15, no. 1, pp. 55–65, 2004.

5. E. Miller and J. Fisher, "ICA using spacings estimates of entropy," *Journal of Machine Learning Research*, vol. 4, pp. 1271–1295, 2003.
6. F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2002.
7. J. Eriksson and K. Koivunen, "Characteristic-function based independent component analysis," *Signal Process*, vol. 83, no. 10, pp. 2195–2208, 2003.
8. A. Chen and P. J. Bickel, "Consistent independent component analysis and prewhitening," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3625–3632, 2005.
9. A. Gretton, R. Herbrich, A. J. Smola, O. Bousquet, and B. Schölkopf, "Kernel methods for measuring independence," *Journal of Machine Learning Research*, vol. 6, pp. 2075–2129, 2005.
10. S. Jegelka and A. Gretton, "Brisk kernel ICA," To be published *Large Scale Kernel Machines* editted by L. Bottou, O. Chapelle, D. DeCoste, and J. Weston. MIT Press, 2007.
11. H. Shen and K. Hüper, "Local convergence analysis of FastICA," in *Lecture Notes in Computer Science, Proceedings of the* 6^{th} *International Conference on Independent Component Analysis and Blind Source Separation (ICA 2006)*, Berlin/Heidelberg, 2006, vol. 3889, pp. 893–900, Springer-Verlag.
12. H. Shen, K. Hüper, and A.-K. Seghouane, "Geometric optimisation and FastICA algorithms," in *Proceedings of the* 17^{th} *International Symposium of Mathematical Theory of Networks and Systems (MTNS 2006)*, Kyoto, Japan, 2006, pp. 1412–1418.
13. H. Shen and K. Hüper, "Newton-like methods for parallel independent component analysis," in *Proceedings of the* 16^{th} *IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2006)*, Maynooth, Ireland, 2006, pp. 283–288.
14. H. Shen, K. Hüper, and A. J. Smola, "Newton-like methods for nonparametric independent component analysis," in *Proceedings of the* 13^{th} *International Conference on Neural Information Processing, Part I, (ICONIP 2006), Hong Kong, China, October 3-6, 2006*, Berlin/Heidelberg, 2006, vol. 4232 of *Lecture Notes in Computer Science*, pp. 1068–1077, Springer-Verlag.
15. A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.
16. J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non gaussian signals," *The IEE Proceedings of F*, vol. 140, no. 6, pp. 363–370, 1993.
17. J.-F. Cardoso, "High-order contrasts for independent component analysis," *Neural Computation*, vol. 11, no. 1, pp. 157–192, 1999.
18. V. Zarzoso and A. K. Nandi, "Blind separation of independent sources for virtually any source probability density function," *IEEE Transactions on Signal Processing*, vol. 47, no. 9, pp. 2419–2432, 1999.
19. K. Hüper, *A Calculus Approach to Matrix Eigenvalue Algorithms*, Habilitation Dissertation, Department of Mathematics, University of Würzburg, Germany, July 2002.
20. M. Kleinsteuber, U. Helmke, and K. Hüper, "Jacobi's algorithm on compact Lie algebras," *SIAM journal on matrix analysis and applications*, vol. 26, no. 1, pp. 42–69, 2004.
21. M. Kleinsteuber, *Jacobi-Type Methods on Semisimple Lie Algebras – A Lie Algebraic Approach to the Symmetric Eigenvalue Problem*, Ph.D. thesis, Bayerische Julius-Maximilians-Universität Würzburg, 2006.
22. A. Kankainen, *Consistent testing of total independence based on empirical characteristic function*, Ph.D. thesis, University of Jyväskylä, 1995.